Scientia Iranica D (2020) 27(3), 1424–1436

SCIENTIA
IRANICA

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
http://scientiairanica.sharif.edu

Recognizing involuntary actions from 3D skeleton data using body states

M. Mokari, H. Mohammadzade*, and B. Ghojogh

Department of Electrical Engineering, Sharif University of Technology, Azadi Ave., Tehran, P.O. Box 11155-4363, Iran.

Received 3 June 2017; received in revised form 13 January 2018; accepted 3 April 2018

KEYWORDS
Human action recognition;
Involuntary action recognition;
Fisher;
Linear Discriminant Analysis (LDA);
Kinect;
3D skeleton data;
Hidden Markov Model (HMM).

Abstract. Human action recognition has been one of the most active fields of research in computer vision over the last years. Two-dimensional action recognition methods face serious challenges such as occlusion and missing the third dimension of data. The development of depth sensors has made it feasible to track positions of human body joints over time. This paper proposes a novel method for action recognition that uses temporal 3D skeletal Kinect data. This method introduces the definition of body states; then, every action is modeled as a sequence of these states. The learning stage uses Fisher Linear Discriminant Analysis (LDA) to construct discriminant feature space for discriminating the body states. Moreover, this paper suggests the use of the Mahalonobis distance as an appropriate distance metric for the classification of the states of involuntary actions. Hidden Markov Model (HMM) is then used to model the temporal transition between the body states in each action. According to the results, this method significantly outperforms other popular methods with a recognition (recall) rate of 88.64% for eight different actions and up to 96.18% for classifying the class of all fall actions versus normal actions.

1. Introduction

Since the last two decades, human action recognition has drawn much attention from researches in computer vision and machine learning fields. In early attempts for action recognition, Red-Green-Blue (RGB) video was used as input of recognition system. Various valuable methods and algorithms were proposed for recognizing actions and activities using RGB data. However, several problems exist in action recognition using RGB frames such as occlusion and different orientations of the camera. The existence of other objects in addition to human bodies and the lack of information of the third dimension can be mentioned as other challenges in this category of methods [1–6]. In order to address these problems, methods for recognizing action from multiple views have been also introduced; however, they are typically very expensive in calculations and are not suitable for real-time recognition [7].

Considering the mentioned problems and introduction of 3D Kinect sensors in the market, researchers have started to work on 3D data for the purpose of action recognition. The Kinect sensor provides both depth and skeleton data in addition to capturing RGB frames. Different methods have been proposed so far for action recognition using either depth or skeleton data.

Action recognition has a variety of different applications. From one point of view, all actions can be categorized in one of the two categories of normal (voluntary) and involuntary actions (see Figure 1). Daily actions, actions for gaming, and interactions between human and computer can be considered as

*. *Corresponding author. Tel.: +98 21 66165927*
E-mail addresses: mozhgan.mokari@ee.sharif.ir (M. Mokari); hoda@sharif.edu (H. Mohammadzade); ghojogh_benyamin@ee.sharif.edu (B. Ghojogh)

doi: 10.24200/sci.2018.20446

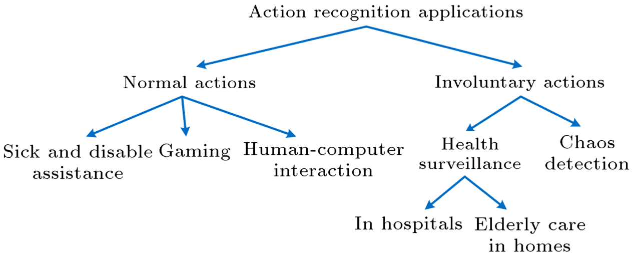


Figure 1. Applications of human action recognition.

normal actions. On the other hand, involuntary actions can occur in different places, such as homes, hospitals, and public places. One of the most frequent involuntary actions is falling which can occur to patients in hospitals. Old people are also subject to dangerous falls, which can reduce serious injuries and fatalities if fall is detected by surveillance systems for elderly care. Another example where proper detection of involuntary actions can prevent problems and chaos is in public places. In these places, involuntary actions such as falling or being thrown can happen as a result of accident or physical fight. In comparison to normal actions, involuntary actions usually have a larger performance variance among various trails and different subjects. This characteristic of involuntary actions is the main challenge of recognizing them. Although the proposed method in this work can be applied to both normal and involuntary actions, its focus is on involuntary actions and tries to handle the mentioned challenge. Figure 2 depicts a human action recognition system used for fall detection. As is seen, it is not possible to train the system using all various types of fall actions over all different subjects. Therefore, the challenge is to recognize any fall action using a limited number of training samples.

This paper proposes a new method for human action recognition, especially for involuntary actions. The main contributions are as follows:

- In contrast to most of the action recognition methods in the literature, this work is not feature-based, but is holistic. In other words, features (such as a histogram of joints as used in [8]) are not extracted from skeleton; the raw features of skeletons are fed to the so-called body state classifier. Consequently, the classifier is responsible for extracting discriminant features. As it is well known in face recognition [9], holistic methods have greater potential for accurate recognition because of using all the information and devolving feature extraction to the classifier. Our experiments verify our better performance in comparison to feature-based methods, such as [8], in both action scenarios of action-vs-action and normal-vs-fall;
- This work properly handles involuntary actions, which are variously distributed in the space of joints, by taking into account the distribution for each body state;
- Different speeds in performing involuntary actions are handled by using Hidden Markov Models (HMM);
- This method can be used for recognizing normal actions as well as involuntary ones;
- Other than outperforming the recognition of each of the various normal and involuntary actions in the dataset, the proposed method achieves a great recognition rate for classifying the class of all involuntary actions versus normal actions. This scenario is particularly important where only the involuntary action detection is important, such as elderly or patient surveillance.

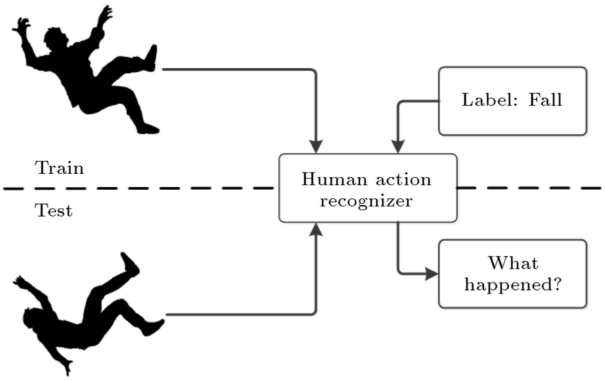


Figure 2. A human action recognition system for fall detection.

This paper is organized as follows. Section 2 reviews related work. Section 3 proposes the main algorithm of the proposed method that includes modeling human body, action recognition using fisher Linear Discriminant Analysis (LDA), and HMM. Section 4 introduces the utilized dataset and experimental results. Finally, Section 5 concludes the article and addresses possible future work.

2. Related work

According to the importance of action recognition and its large amount of applications, many different methods have been proposed in this field. In [10], Peng et al. described different kinds of Bag of Visual Words model (BoVW) methods and investigated the effect of each of them on action recognition. These factors included feature extraction, feature preprocessing, codebook generation, feature encoding, pooling, normalization, and fusing these descriptors.

Liu et al. [11] employed Genetic Programming (GP) on spatio-temporal motion features for action recognition. Features were extracted from both color and optical flow sequences. Wang et al. [12] used homography for cancellation of camera motions from trajectories and optical flows. SURF descriptors and dense optical flows were employed with RANSAC to estimate this homography. Then, a motion-based histogram of optical flows and motion-based histogram descriptors were used for action recognition.

Facing some challenges such as the coverage of some body parts by others and introducing 3D methods, encouraged researchers to use depth map. Li et al. [13] recognized human's action by sampling the 3D points of the depth image and creating an action graph. In this method, they modeled the position of human body by projecting the contour of body shape onto different planes and sampling them. Then, the state of human body was modeled with these bags of 3D points. The states were considered as nodes of a graph, modeling the action. Although this method is not robust to the constant change of viewing angle and human's body scale, it recognized 90% of actions, and the error was halved compared to 2D methods. Rahmani et al. [14] used histogram of oriented principal components descriptor on point clouds for cross-view action recognition.

Zhao et al. [15] classified human's actions by utilizing information of RGB and depth image. They obtained spatiotemporal interest points from RGB image and used a combined descriptor of RGB and depth images.

Liu et al. [16] encoded spatio-temporal information of skeleton joints in depth sequences into color images. In this regard, 5D space of (x, y, z, f, n) was expressed as a 2D coordinate space and a 3D color space, where f and n denote time and joint labels, respectively. A convolutional neural network was used to extract more discriminative deep features. These features were used for action recognition.

Rahmani and Mian [17] transferred human poses to a view-invariant high-level space and recognized action in depth image by using a deep convolutional neural network. Their method obtained appropriate results in multi-view datasets. In [18], Zhang et al. used 3D Histograms of Texture (3DHoTs) from depth maps. The 3DHoTs were formed by characterizing the salient information of action. In their method, action was represented by texture features. The classification of actions was done by the multi-class boosting classifier.

Chen et al. [19] projected depth videos onto three orthogonal Cartesian planes. The absolute difference between two consecutive projections was accumulated, creating Depth Motion Maps (DMMs). Then, action recognition was performed by distance-weighted Tikhonov matrix with an I2-regularized classifier. Chen et al. [20] proposed a Local Binary Patterns (LBP) descriptor which is invariant to shape and speed for action recognition in depth videos. They partitioned DMMs and extracted LBP for action recognition. Liang et al. [21] applied DMMs-based gradient local autocorrelations features of depth videos to capture the shape information of sub-actions. They proposed Locality-constrained affine subspace coding to encode the extracted features. This method had competitive results with less complexity.

By developing Kinect sensors and related software for tracking humans in images and detecting positions of body joints in 3D space, several methods were proposed to recognize action using this information. One of these methods introduced Cov3DJ descriptor [22], which separated different action classes by finding covariance matrix of positions of the joints during the action and used Support Vector Machine (SVM) for classification.

Reddy et al. [23] recognized action by considering mean, minimum, and maximum of position of joints as features and compared them to features obtained by using Principle Component Analysis (PCA) on position of joints. Likewise, Martínez-Zarzuela et al. [24] tried to recognize actions by taking a sequence of positions of joints as a signal and extracting the five first Fast Fourier Transform (FFT) components as a feature vector fed into a neural network. However, this method did not perform very well for complex actions involving different body parts.

As different actions involve different joints, Anjum et al. [25] selected important and effective joints at the training level, according to the type of action. In their method, each action was determined by three joints. Results showed that this method performed better with less information; however, joints should be selected in training for each action. Therefore, extending this algorithm for new actions is time consuming and expensive.

Liu et al. [26] used a tree-structure-based traversal method for 3D skeleton data and extended RNN-based learning method to spatio-temporal domain. In this way, they could analyze the hidden sources of information in actions. Ke et al. [27] transformed skeleton sequences into clips consisting of spatio-temporal features. They used deep convolutional neural networks to learn long-term temporal information. Multi-task learning network was used to incorporate spatial structural information for action recognition.

In [28], Shahroudy et al. described actions by partitioning kinetics of body parts. They used a sparse set of body parts to model actions as a combination of multimodal features. Dynamics and appearance of parts were represented by a heterogeneous set of depth and skeleton-based features. Huynh et al. [29] proposed a new method more robust to human scale and changes of position. They categorized joints into three classes of stable, active, and highly active joints and utilized angles of 10 important joints and vectors connecting moving joints to stable joints. Their method performed better than a similar method, which uses only raw positions of joints.

Luvizon et al. [30] selected subgroups of joints by vector of locally aggregated descriptors algorithm. Classification accuracy was improved by the non-parametric KNN classifier with large margin nearest neighbor. Amor et al. [31] used skeleton shapes as trajectories on Kendall's shape manifold to represent special dynamical skeletons.

Xia et al. [8] used middle and side hip joints to extract a histogram of positions of other joints to be used as a feature vector. They reduced the dimension of the feature vector using LDA and used K-means method to cluster the feature vectors. Each cluster constituted a visual word. Each action was determined as a time sequence of these visual words and modeled by HMM. Results showed that this method partially overcame challenges such as different lengths of actions and the same action done in different ways and view angles.

Papadopoulos et al. [32] obtained the orientation of body using the positions of shoulders and hip joints and, thereby, extracted orthogonal basis vectors for each frame. A new space was then constructed for every person according to its orientation of body. According to these vectors and the new space, the spherical angles of joints were used instead of positions of joints. The use of angles, instead of position of joints, made the method more robust against human's body scale and changes in the shape of body. This method also used the energy function to overcome the challenge of the same actions done by opposite hands or feet.

Although there are many proposed methods for action recognition, many problems and challenges still remain unsolved. This paper tries to tackle some of them such as different distributions of actions in statistical feature space, especially for involuntary actions.

3. Methodology

In order to recognize actions, in the first step, the actions should be modeled appropriately. Modeling actions depends on various facts such as application, types of actions, and method of classification. One of the most important applications of action recognition is online recognition where the recognition should be performed in real time. This article considers this type of recognition as its objective. In this category, the action should be modeled so that the model can be updated during completion of action and, finally, the type of the performed action can be recognized. Therefore, in this article, each action is supposed to be a sequence composed of several body states.

In the next step, positions of joints in the 3D space are utilized in order to model the state of body. The positions of joints are prepared by the output of the Kinect sensor. The skeleton consists of several joints, which are 25 joints for the dataset used for dealing with the experiments in this paper. A number of these joints are, however, very close to each other without any important difference in movements; therefore, their information is almost redundant. With respect to the actions addressed in this paper, merely 12 important joints, including right and left ankles, right and left knees, right and left wrists, right and left shoulders, head, middle spine, hip and spine shoulder, are selected out of the skeleton. The position of spine base (hip) and right and left shoulders are used for alignment in order to correctly describe the state of body in different persons and runs. The selected joints and, also, the joints required for alignment are shown in Figure 3. State modeling, including skeleton alignment and state classification, is detailed in the following.

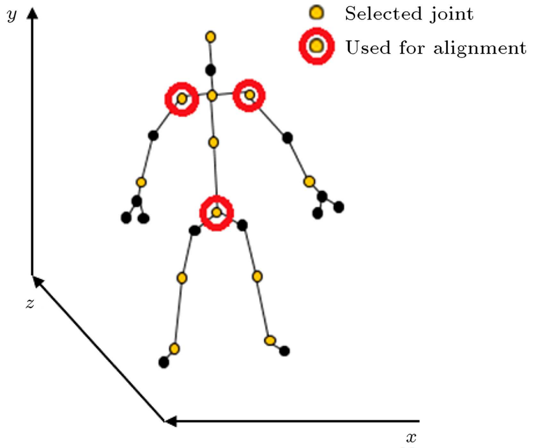


Figure 3. Selected joints out of available joints in the skeletal data. The joints used for alignment are also shown.

3.1. Modeling state of body

In order to model and describe the state of body, a proper descriptor should be created. This descriptor models the action as a time sequence of states and tries to recognize the action. The body states are determined as follows. According to the nature of every action, the main body states, of which the action is composed, are conjectured and, then, are manually selected and sampled out of the training sequences of frames. Notice that this manual selection is done merely in the training phase, while, in the test phase, each input frame is automatically classified by the classifier of body states.

1. *Aligning skeleton:* Different locations and orientations of body in the frames create the need to align the skeleton. As already mentioned, 12 joints positions are used in 3D space in order to describe the state of body. In order to cancel the location of body skeleton, the position of hip joint is subtracted from the position of all joints. This is performed for every frame in the sequence.

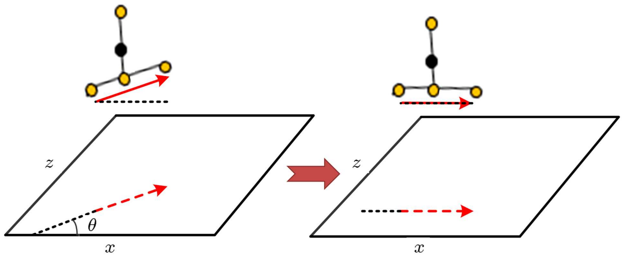


Figure 4. Alignment of skeleton using the left and right shoulders to cancel the orientation of skeleton.

Moreover, different orientations of skeleton or camera make recognizing similar states difficult and even wrong. Thus, in order to cancel different orientations of body skeletons, the body rotates around y-axis making the projection of the vector that connects the left and right shoulders onto xz plane parallel to x-axis. Through this rotation, the skeleton directly faces the camera. This procedure is illustrated in Figure 4. The axes can be seen in Figure 3. In the literature, the alignment of skeleton is often performed; however, the methods or the joints used for that might differ. For example, in [8], left and right hip joints are utilized rather than shoulder joints for alignment;

2. *Creating feature vector:* To determine the state of body in each frame, proper feature vectors are required. Three joints out of the 12 joints are used for alignment and the remaining nine joints are used to create the feature vectors. If (x_m, y_m, z_m) denote the coordinates of the mth joint $(m = \{1, \cdots, 9\})$, the raw feature vectors are obtained as $[x_1, \cdots, x_9, y_1, \cdots, y_9, z_1, \cdots, z_9]^T$. Fisher LDA [33,34] is utilized for extracting discriminant features from the raw feature vectors. In Fisher LDA method, the dimension of the feature vector is reduced to $C-1$, where C is the number of states. In LDA, the within (S_w) and between-class (S_b) scatter matrices are:

$$S_w = \sum_{i=1}^{C} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T, \tag{1}$$

$$S_b = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T, \tag{2}$$

in order to minimize the within class covariance and maximize the between class covariance [34,35], where μ_i and μ denote the mean of the ith state and the mean of class means, respectively. The Fisher projection space is created by the eigenvectors of $S_w^{-1} S_b$. By its projection into this space, feature vector F for an input skeleton state is obtained.

After projection onto Fisher space, the obtained feature vectors are located relative to each other such that those relating to similar and different states, respectively, fall close and apart. By this fact, recognition of states becomes available.

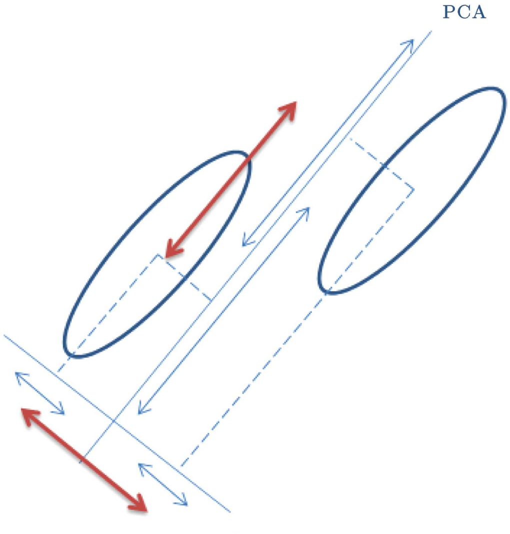


Figure 5. An example of Fisher and Principle Component Analysis (PCA) directions.

There are also other methods for feature reduction which can be used for classification. One of the most popular methods of this category is PCA [34,35]. However, PCA method cannot always classify the data as well as LDA does. As an example, suppose that the distribution of classes is similar to that depicted in Figure 5. In this example, the Fisher LDA direction is perpendicular to the direction of PCA. As is obvious in this figure, Fisher LDA tries to minimize within-class variance while maximizing between-class variance in order to classify them.

The resulting feature vectors are used for training and testing the state of body. The action will be defined as a time sequence of multiple specific states. The state of body is recognized in the test phase by finding the minimum distance as described in the following section.

3. *Finding the minimum distance*: In every frame denoted as f, the state of body should be recognized. For achieving this goal, the distances between feature vector F of this frame and the means of the feature vectors of all states are found. The minimum distance determines the state of frame f. If $\widetilde{F}_i$ denotes the mean of feature vectors of the ith class, the state is found as:

$$\text{state}(f) = \arg\min_i d\left(F, \widetilde{F}_i\right), \tag{3}$$

where d is the distance measurement function which can be one of the two followings:

- *Euclidean distance*: One of the most popular methods for calculating the distance of two vectors is Euclidean distance, which is used as one of the distance methods in this article. The function of Euclidean distance can be formulated as:

$$d\left(F,\widetilde{F}_i\right)=\sqrt{\sum_j\left(F_j-\widetilde{F}_{ij}\right)^2}, \qquad (4)$$

where F_j and $\widetilde{F}_{ij}$ are the jth components of F and $\widetilde{F}_i$, respectively;

- *Mahalanobis distance*: As the minimum distance from the means of states is used for recognizing the state, using a proper distance has much important influence on the accuracy of recognition. Therefore, the distribution of final feature vectors in the feature space should be considered, and the distance measurement should be defined accordingly.

  If body states are categorized into C classes, the dimension of the final feature (Fisher) vectors will be $C-1$. As the dimension of the feature vectors might be high, their distribution in each class cannot be directly visualized for direct analysis. However, the distribution of feature vectors can be analyzed in higher dimensions by calculating their covariance matrices. The first twodirections of Fisher space are used here for illustrating the distribution of each of the eight body states defined for the TST dataset [38,39], which are discussed in more details in Section 4. Figure 6 illustrates the training samples projected onto the space constructed by the first two Fisher directions. As shown in this figure, distribution of feature vectors for each state is different in the two directions.

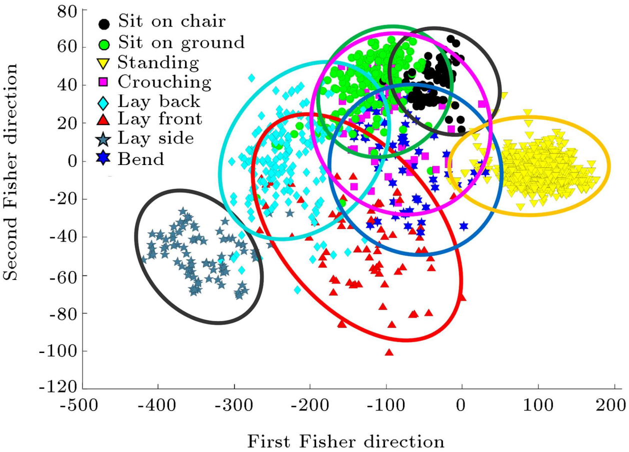


Figure 6. Projection of samples of states onto Fisher space. As can be seen, the states have different distributions.

The more differently people perform an action containing a state, the wider the distribution for the state would be. The more widely distributed states are usually those during the completion of an involuntary action. For instance, as shown in Figure 6, after projection on constructed Fisher space, states related to normal actions such as standing and sit states are less distributed than the states occurred in involuntary actions, such as lay front and lay back. In order to handle the challenge of different distributions of projected states, a distance measurement function other than Euclidean one should be used which considers the distributions.

Mahalanobis distance considers variances of distributions in its calculation, which is calculated as:

$$d\left(F,\widetilde{F}_i\right)=\sqrt{\left(F-\widetilde{F}_i\right)^T S^{-1}\left(F-\widetilde{F}_i\right)}, \qquad (5)$$

where S denotes the covariance matrix of the feature vectors of the class to which the distance is calculated.

As is obvious in Eq. (5), covariance matrix S acts as a weighting matrix for each class according to its distribution. That is, the importance of distance in a particular dimension is considered in calculating the distances. In other words, the distance in a direction with smaller variance is less valuable, yielding S^{-1} in the equation.

Mahalanobis distance is actually an extension to the standard deviation from the mean in the multi-dimensional space. The experiments reported in the following sections show the outperformance of this distance in comparison with Euclidean distance.

3.2. *Classyfing actions using HMM*

As previously mentioned, every action can be modeled as a sequence of consequent states. After recognizing states of body using Fisher LDA, HMM is utilized in this work to classify actions.

Every action is modeled using a separate HMM. Each HMM has a number of hidden states with specific transition probabilities between them. For instance, a three-state HMM and its transition probabilities are illustrated in Figure 7 [36]. Every hidden state has specific emission probabilities for emitting body states. The transition and emission probabilities of each HMM are estimated by the well-known Baum-Welch expectation maximization algorithm [37] using the training observations, i.e., sequences of body states. This algorithm starts with initial assumptions for all of the parameters of the model (i.e., transition and emission probabilities) and then updates the param-

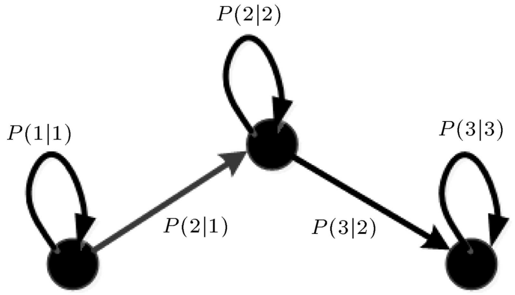


Figure 7. A three-state Hidden Markov Model (HMM) model.

eters using corresponding expectation maximization equations iteratively until convergence.

In order to decrease computational cost of the algorithm, the frame per second rate has been reduced by down sampling. Uniform down sampling with the rate of 20 is used which was shown appropriate according to our experiments (note that the sampling rate of the Kinect V2 sensor is known to be 30 frames per second (fps) in normal lighting conditions and 15 fps in poor lighting conditions. According to the corresponding RGB images of the dataset, the samples in this dataset should have been captured in normal lighting condition, and hence, the original sampling rate for this dataset must be 30 fps). After constructing a HMM for each action, an unknown sequence is recognized by feeding it to each HMM. After feeding the test sequence of frames to all trained HMMs, every HMM outputs a probability of occurrence for that sequence. The maximum probability determines the action of that sequence.

To obtain an insight, note that every period of repetitions of a body state can be roughly associated to a HMM state. For example, when having three-state HMMs for classifying actions, the actions sit, grasp, and end up sit are mostly made of the sequences:

- $\{\overline{\text{standing}}, \overline{\text{crouching}}, \overline{\text{sit on chair}}\}$,
- $\{\overline{\text{standing}}, \overline{\text{bend}}, \overline{\text{standing}}\}$,
- $\{\overline{\text{standing}}, \overline{\text{crouching}}, \overline{\text{sit on ground}}\}$

where $\overline{\text{body state}}$ denotes a sub-sequence of repetitions of the body state (more details about how body states are defined are discussed in Section 4). Moreover, in each sub-sequence, the number of repetitions of the corresponding body state can be different across subjects and different trials.

For each action, the sequences that are used for training HMM are adjusted to have the same lengths (number of body states). This equalization is performed by manually repeating the last state done by the person so that the total number of states of all actions becomes equal. It is important to note that this equalization does not compensate for different lengths and speeds of actions performed by different people or over different trials.

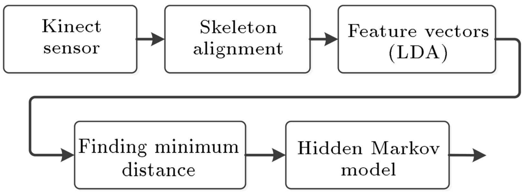


Figure 8. The overall structure of the proposed framework.

The advantage of HMM, in this work, is that it considers solely the dynamic of sequence and is not sensitive to various paces and lengths of actions. For instance, there exist sequences of lengths 75 frames upto 463 frames with different speeds of actions in TST fall dataset [38,39], and these sequences have been successfully handled and recognized by this method.

The overall structure of the proposed framework is summarized in Figure 8.

4. Experimental results

To examine the proposed method, TST Fall Detection dataset [38] is used. The details of this dataset are explained in next section followed by the explanation on how the actions are modeled in this dataset. In the end, the results of the experiments are presented.

4.1. Dataset

TST Fall Detection dataset [38,39] is used for verifying the effectiveness of this method. There are two main categories of actions in this dataset, i.e., daily living activities and fall actions. Eleven different persons perform every action for three times. The daily living activities are sit, lay, grasp, and walk; the fall actions are falling front, back, side, and end up sit.

This dataset has prepared information of 3D position of joints and depth data obtained by the Kinect sensor V2, which is more accurate than previous Kinect sensors. Only the skeletal data of this dataset are used in this work for experiments.

As previously mentioned, one of the important goals of human action recognition is surveillance application, especially for controlling the elderly or patients. The main goal of detecting involuntary actions and improvements of Kinect V2 encouraged this work to use the mentioned dataset. Unlike other datasets, involuntary actions, such as falling down, exist sufficiently in this dataset, making this database challenging.

As fall actions are performed involuntarily, different states and conditions from normal actions appear for different people. Therefore, existing action recognition methods may not necessarily perform as well for fall actions. Moreover, a number of methods have been proposed to recognize fall actions, which concentrate on

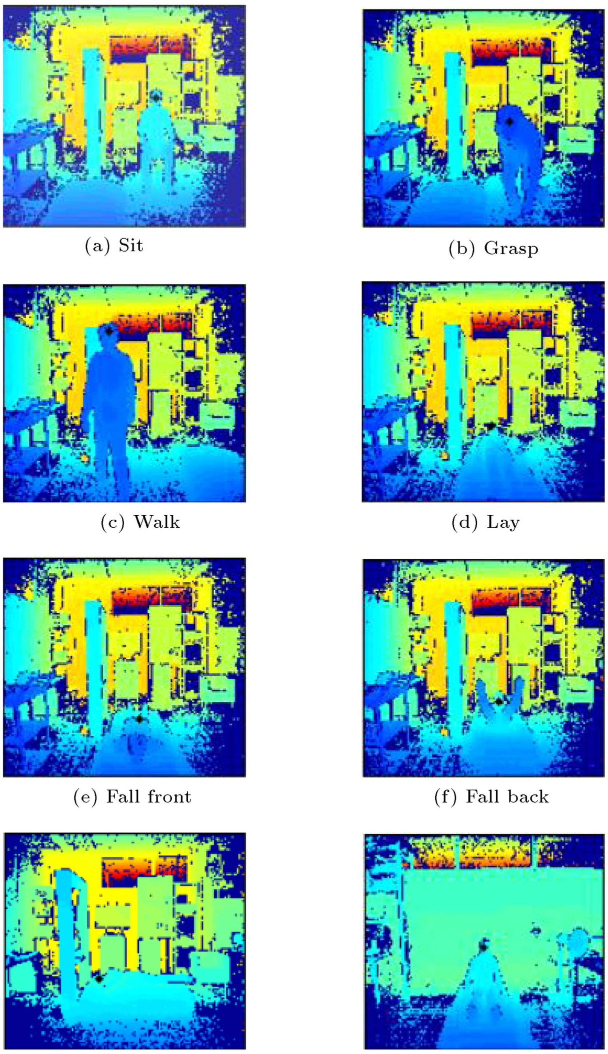


Figure 9. An example of actions in TST dataset [38,39].

using features such as speed and acceleration recorded by accelerometer sensors. These features are not able to effectively discriminate the normal actions from each other nor involuntary actions from each other; thus, they do not help recognize the actions in general. Therefore, the main challenge here is to develop a method that can detect and analyze both of the normal and involuntary actions and, also, recognize them from each other.

Several samples of depth images of actions in TST dataset are shown in Figure 9.

4.2. Recognition of states

In the dataset, only the actions are labeled; therefore, labeling states should be performed manually. According to the actions, eight different states are chosen and labeled to be used to train and test the state classification module. The chosen states should include the main states of actions in the dataset and should not contain unnecessary states that are close to other states. The chosen states are standing, crouching, lay back, lay front, lay side, bend, sit on chair, and sit on ground. An example of each state is shown in Figure 10.

Table 1. Correctness rate of recognizing state of body.

State	Euclidean	Mahalanobis
Standing	99.38%	94.26%
Crouching	50.00%	70.00%
Lay back	80.71%	81.22%
Lay front	67.50%	85.00%
Lay side	88.89%	82.22%
Bend	62.90%	90.32%
Sit on chair	86.87%	69.70%
Sit on ground	72.15%	79.91%
Total	**76.03%**	**81.57%**

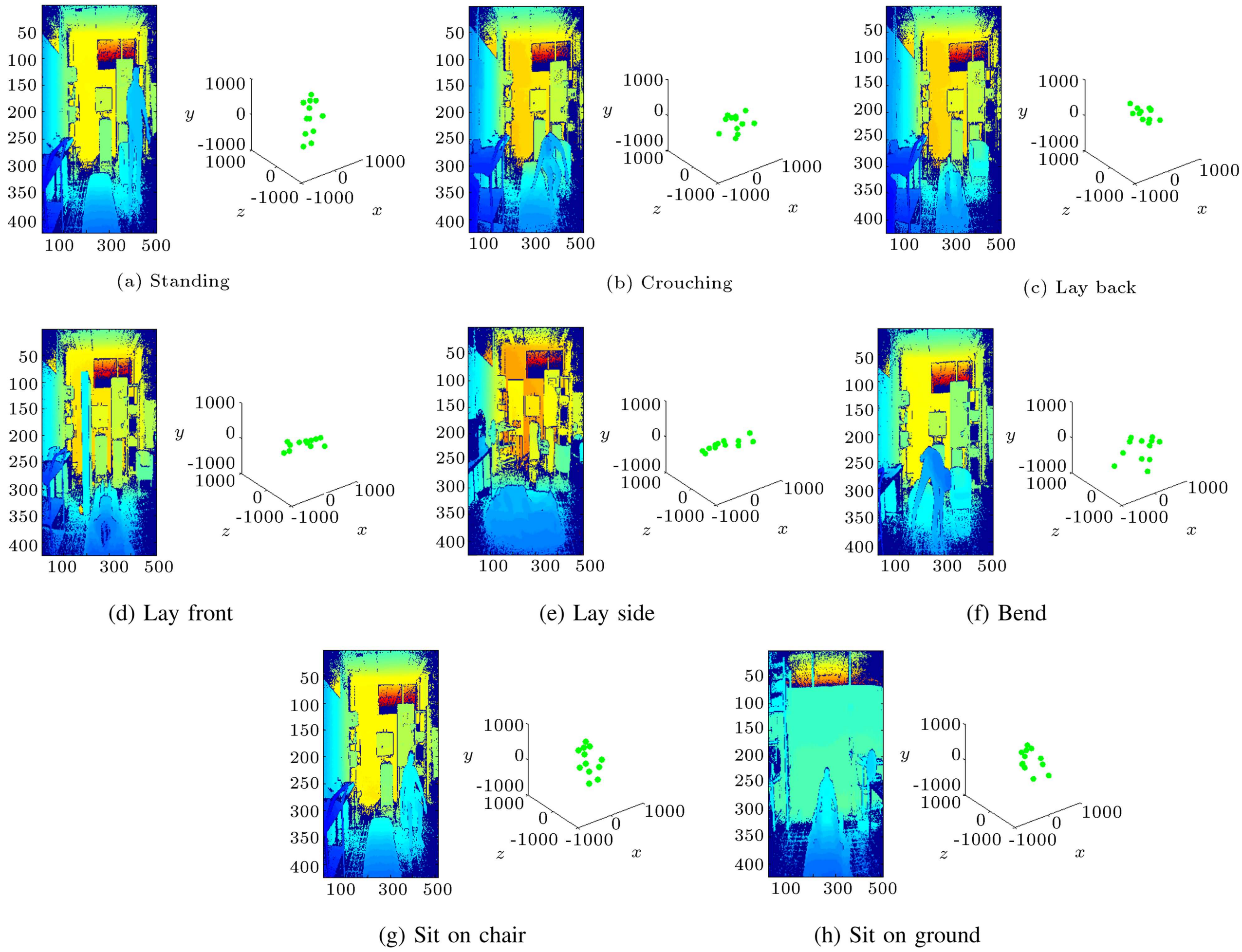
The "leave one subject out" cross-validation is used for the experiments. In each iteration, the entire samples of a person are considered as test samples, and the samples of other subjects are used for training system. This type of cross-validation is fairly difficult, because the system does not see any sample from the test subject in training phase. The state recognition experiment is repeated using both of the distance methods, and the results are listed in Table 1. Note that all the rates reported in this paper are recall rates:

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \tag{6}$$

unless the type of rate is mentioned.

Table 1 shows that the Mahalanobis distance outperforms the Euclidean distance in general. As was expected, the recognition rates of crouching, lay front, and bend have improved significantly using Mahalanobis distance. The reason is that the variances of training data for these states are huge, and this fact is not taken into account when Euclidean distance is used.

It is worth noting that by using the Mahalanobis distance, the recognition rate of bend state has improved at the cost of reducing the recognition rate of standing state. A careful consideration of Figure 6 reveals that there exists an overlapping region of distributions between the two states. Euclidean distance, which does not consider the distribution of classes, mostly recognizes the overlapping region as the standing state. On the other hand, the Mahalanobis distance mostly recognizes this region as the bend state, because the variance of standing state is much less than bend. This fact can also be seen from the confusion matrices of states for both distances, which are depicted in Figure 11.



(a) Standing

(b) Crouching

(c) Lay back

(d) Lay front

(e) Lay side

(f) Bend

(g) Sit on chair

(h) Sit on ground

Figure 10. An example of the selected states.

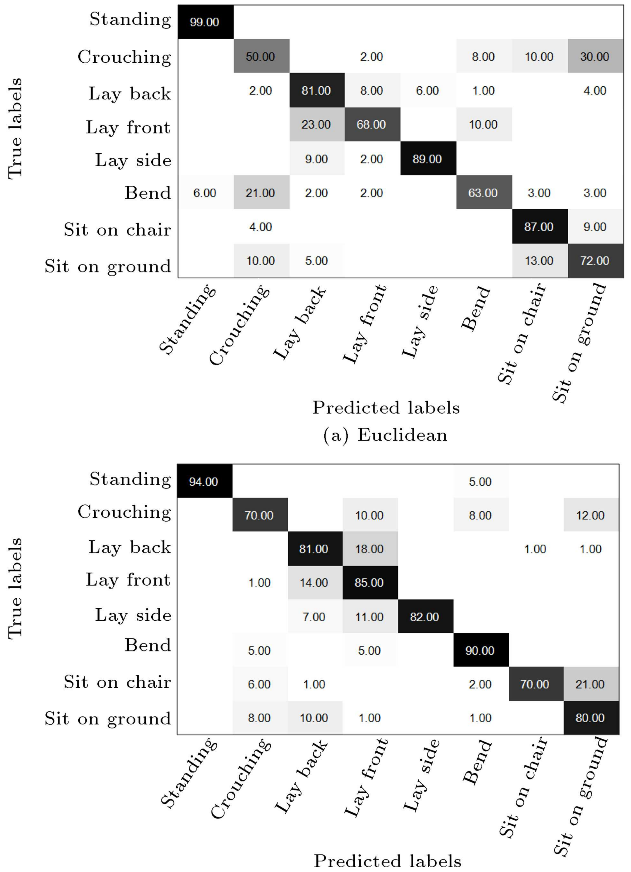
4.3. Action recognition and comparison

In the last step, an action is represented as a sequence of states. Each state in this sequence is recognized by projecting into the LDA space and utilizing a distance measure. Then, the probability of each HMM (note that there is an HMM for each specific action) generating the input sequence of states is calculated, and the maximum probability determines the recognized action. The number of hidden states in HMMs (note that hidden states are different from body states) affects the recognition performance. Therefore, different hidden states were tested for HMMs in this work and were compared to each other. Results of three different numbers of hidden states for HMMs are reported in Table 2. The experiments of this table are performed with Mahalanobis distance. As was expected, according to the nature of states and actions in the TST Fall dataset [38,39], HMMs with three hidden states perform better; hence, the number of hidden states for HMMs is considered to be three in this work. It is worth noting that the combination of optimum number of hidden states for each action was also considered; however, the experiments showed that the use of a constant number of hidden states for all HMMs results in better performance.

Table 2. Effects of the number of states of HMM on the recognition rate.

Action	**2 states**	**3 states**	**4 states**
Sit	87.88%	90.91%	90.91%
Grasp	90.91%	90.91%	87.88%
Walk	93.94%	93.94%	93.94%
Lay	84.85%	96.97%	90.91%
Fall front	84.85%	81.82%	81.82%
Fall back	84.85%	84.85%	78.79%
Fall side	81.82%	81.82%	81.82%
End up sit	84.85%	87.88%	84.85%
Total	86.74%	**88.64%**	86.36%

In this article, the proposed method is compared with the method of Xia et al. [8] which has received considerable attention in literature [40–43] and has



(a) Euclidean

(b) Mahalanobis

Figure 11. Confusion matrix of states.

been used for comparison in very recent methods [44–48]. Note that all the above methods have been experimented with the datasets that have been created using an older version of Kinect sensor with no involuntary actions.

For implementing this method [8] and fairly comparing it with the proposed method using the TST dataset, several necessary adjustments were done to its settings. First, for LDA, the states were labeled in the same way as in the proposed method. Second, the number of hidden states for HMMs was chosen to be three, according to the actions of the dataset. Third, the best number of clusters for histogram was experimented to be eight, which conformed to the number of classes of states in the proposed method.

Results are listed in Table 3. The proposed method using both of the distance methods is compared with the method of Xia et al. [8]. Results reveal that, in all actions, the proposed method using each of the two distance measures outperforms the method [8]. Although the method [8] has utilized LDA and clustering methods to prepare data for training HMM, it has made several states very close to each other by using a histogram concept, which has increased the error. As an example, in fall actions, the angular positions of joints are much similar, and their differences are ignored when using the histogram.

Table 3. Comparison of results of our method and method [8] for TST dataset.

Action	Euclidean	Mahalanobis	[8]
Sit	84.85%	90.91%	81.82%
Grasp	96.97%	90.91%	84.85%
Walk	100%	93.94%	90.91%
Lay	75.76%	96.97%	90.91%
Fall front	54.54%	81.82%	48.49%
Fall back	69.70%	84.85%	66.67%
Fall side	81.82%	81.82%	69.70%
End up sit	69.70%	87.88%	33.33%
Total	79.16%	**88.64%**	70.83%

Using Mahalanobis distance has significantly enhanced the performance, especially in fall actions. In other words, improving the performance of recognizing difficult involuntary states, such as crouching and lay front, has improved the total recognition rate. As mentioned before, the main reason for this fact is that the intrinsic variance of states is considered in the Mahalanobis distance.

The confusion matrix of actions is reported in Figure 12. This matrix shows that the actions that are similar to each other are sometimes confused and wrongly recognized. Actions, such as falling on front, side, and back, are sometimes confused with each other, because their distributions (and, thus, their behavior) are similar and wider than others, as is obvious in Figure 6. In some scenarios such as anomaly detection in actions, this confusion between subgroup actions might not matter. Hence, another experiment was performed considering all fall and normal actions as two different high-level groups. In this scenario, the recognition rate improves from 88.64% to 96.18%. In addition, as can be seen in Table 4, the false alarm rate has also been significantly reduced. This result indicates that the possibility of wrongly recognizing a normal action as fall action is considerably low.

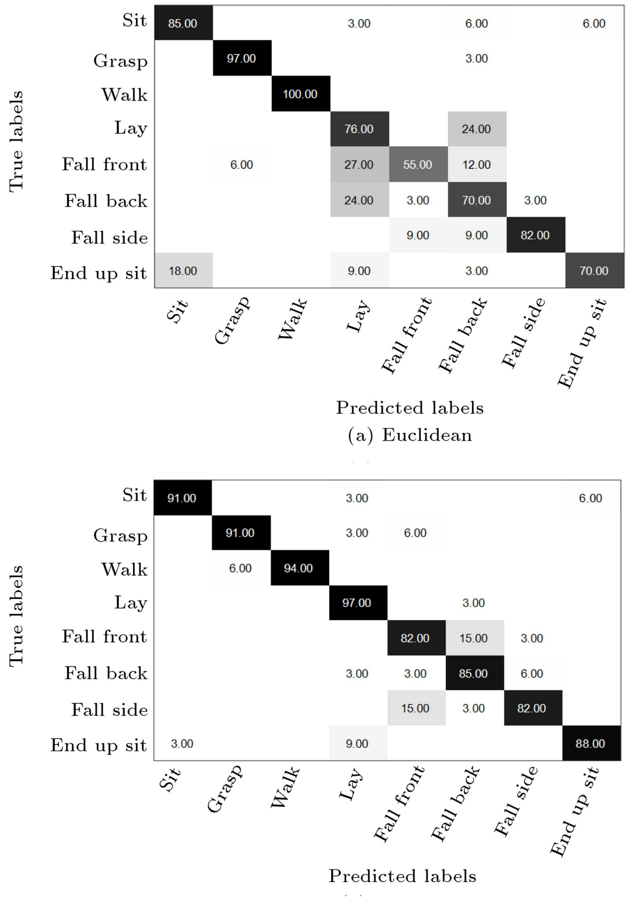
5. Conclusion and future work

5.1. Conclusion

A new action recognition method was proposed in this paper, which is particularly useful for recognizing the actions with some sort of complexities such as various types of falling action. Since this method uses feature vectors with low dimension and does not have big computational overhead, it can be used in real-time purposes. Experiments showed that this method outperformed other methods, especially in scenarios where normal and involuntary actions were mixed up.

Table 4. Comparison of results, considering all abnormal.

	Euclidean	**Mahalanobis**	[8]
Recognition rate (true positive rate)	78.78%	**96.18%**	77.27%
Specificity rate (true negative rate)	90.15%	**96.21%**	90.90%
False alarm rate (false positive rate)	9.15%	**3.78%**	9.09%



(a) Euclidean

(b) Mahalanobis

Figure 12. Confusion matrix of actions.

In the proposed method, a feature vector was created for representing the state of body in each frame using the Kinect data. The state of body was then recognized in the corresponding discriminative Fisher subspace. Finally, actions are classified and recognized by feeding the sequence of recognized states of body to HMMs. Because of using Hidden Markov Model (HMM), this method is robust to different paces and lengths of actions. Moreover, the Mahalanobis distance is utilized for considering the wider distribution of involuntary body states in order to enhance the recognition rate.

5.2. Potential future work

Data were preprocessed by skeleton alignment to make the algorithm robust against the orientation of camera. As for the future work, the angles between the joints can be used instead of their positions in order to obtain more robustness. In addition, recognizing more complex and longer actions can be considered as future work.

Moreover, manual selection/sampling of body states limits the scalability of the system. Automatic selection of body states in an approach similar to [49], which automatically finds elementary states of higher level actions, can also be considered as future work.

Another possible limitation of the proposed method is that canceling the motion of body by alignment, which is necessary for the proposed method, omits the motion information. This cancellation might cause difficulties for recognizing actions with close body states, except for different motions. Handling this issue can be considered as another potential future work.

Acknowledgment

This work was supported by a grant from Iran National Science Foundation (INSF).

References

1. Yao, A., Gall, J., and Gool, L.V. "Coupled action recognition and pose estimation from multiple views", *International Journal of Computer Vision*, **100**(1), pp. 16–37 (2012).
2. Guo, K., Ishwar, P., and Konrad, J. "Action recognition from video using feature covariance matrices", *IEEE Transactions on Image Processing*, **22**(6), pp. 2479–2494 (2013).
3. Wang, H., Kläser, A., Schmid, C., and Liu, C.L. "Dense trajectories and motion boundary descriptors for action recognition", *International Journal of Computer Vision*, **103**(1), pp. 60–79 (2013).
4. Liu, J., Luo, J., and Shah, M. "Recognizing realistic actions from videos in the wild", In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1996–2003 (2009).
5. Wang, H., Klaser, A., Schmid, C., and Liu, C.L. "Action recognition by dense trajectories", In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3169–3176 (2011).
6. Niebles, J.C., Wang, H., and Fei-Fei, L. "Unsupervised learning of human action categories using spatial-temporal words", *International Journal of Computer Vision*, **79**(3), pp. 299–318 (2008).
7. Holte, M.B., Tran, C., Trivedi, M.M., and Moeslund, T.B. "Human action recognition using multiple views:

a comparative perspective on recent developments", In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, ACM*, pp. 47–52 (2011).

8. Xia, L., Chen, C.C., and Aggarwal, J.K. "View invariant human action recognition using histograms of 3D joints", In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 20–27 (2012).

9. Zhao, W., Chellappa, R., Phillips, P.J., and Rosenfeld, A. "Face recognition: A literature survey", *ACM Computing Surveys (CSUR)*, **35**(4), pp. 399–458 (2003).

10. Peng, X., Wang, L., Wang, X., and Qiao, Y. "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice", *Computer Vision and Image Understanding*, **150**, pp. 109–125 (2016).

11. Liu, L, Shao, L., Li, X., and Lu, K. "Learning spatio-temporal representations for action recognition: A genetic programming approach", *IEEE Transactions on Cybernetics*, **46**(1), pp. 158–170 (2016).

12. Wang, H., Oneata, D., Verbeek, J., and Schmid, C. "A robust and efficient video representation for action recognition", *International Journal of Computer Vision*, **119**(3), pp. 219–238 (2016).

13. Li, W., Zhang, Z., and Liu, Z. "Action recognition based on a bag of 3d points", In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 9–14 (2010).

14. Rahmani, H., Mahmood, A., Huynh, D., and Mian, A. "Histogram of oriented principal components for cross-view action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(12), pp. 2430–2443 (2016).

15. Zhao, Y., Liu, Z., Yang, L., and Cheng, H. "Combing rgb and depth map features for human activity recognition", In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, IEEE*, pp. 1–4 (2012).

16. Liu, M., Chen, C., Meng, F., and Liu, H. "3d action recognition using multi-temporal skeleton visualization", In *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*, pp. 623–626 (2017).

17. Rahmani, H. and Mian, A. "3d action recognition from novel viewpoints", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1506–1515 (2016).

18. Zhang, B., Yang, Y., Chen, C., Yang, L., Han, J., and Shao, L. "Action recognition using 3d histograms of texture and a multi-class boosting classifier", *IEEE Transactions on Image Processing*, **26**(10), pp. 4648–4660 (2017).

19. Chen, C., Liu, K., and Kehtarnavaz, N. "Real-time human action recognition based on depth motion maps", *Journal of Real-Time Image Processing*, **12**(1), pp. 155–163 (2016).

20. Chen, C., Liu, M., Zhang, B., Han, J., Jiang, J., and Liu, H. "3D action recognition using multi-temporal depth motion maps and fisher vector", In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp. 3331–3337 (2016).

21. Liang, C., Chen, E., Qi, L., and Guan, L. "3d action recognition using depth-based feature and locality-constrained affine subspace coding", In *Multimedia (ISM), 2016 IEEE International Symposium on*, pp. 261–266 (2016).

22. Hussein, M.E., Torki, M., Gowayyed, M.A., and El-Saban, M. "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations", In *IJCAI*, **13**, pp. 2466–2472 (2013).

23. Reddy, V.R. and Chattopadhyay, T. "Human activity recognition from kinect captured data using stick model", In *International Conference on Human-Computer Interaction, Springer*, pp. 305–315 (2014).

24. Martínez-Zarzuela, M., Díaz-Pernas, F.J., Tejeros-de-Pablos, A., González-Ortega, D., and Antón-Rodríguez, M. "Action recognition system based on human body tracking with depth images", *Advances in Computer Science: An International Journal*, **3**(1), pp. 115–123 (2014).

25. Anjum, M.L., Ahmad, O., Rosa, S., Yin, J., and Bona, B. "Skeleton tracking based complex human activity recognition using kinect camera", In *International Conference on Social Robotics*, Springer, pp. 23–33 (2014).

26. Liu, J., Shahroudy, A., Xu, D., and Wang, G. "Spatio-temporal lstm with trust gates for 3d human action recognition", In *European Conference on Computer Vision*, Springer, pp. 816–833 (2016).

27. Ke, Q., Bennamoun, M., An, S., Sohel, F., and Boussaid, F. "A new representation of skeleton sequences for 3d action recognition", arXiv preprint arXiv:1703.03492 (2017).

28. Shahroudy, A., Ng, T.T., Yang, Q., and Wang, G. "Multimodal multipart learning for action recognition in depth videos", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(10), pp. 2123–2129 (2016).

29. Huynh, L., Ho, T., Tran, Q., Dinh, T.B., and Dinh, T. "Robust classification of human actions from 3d data", In *Signal Processing and Information Technology (ISSPIT), 2012 IEEE International Symposium on*, pp. 263–268 (2012).

30. Luvizon, D.C., Tabia, H., and Picard, D. "Learning features combination for human action recognition from skeleton sequences", *Pattern Recognition Letters* (2017).

31. Amor, B.B., Su, J., and Srivastava, A. "Action recognition using rateinvariant analysis of skeletal shape trajectories", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(1), pp. 1–13 (2016).

32. Papadopoulos, G.T., Axenopoulos, A., and Daras, P. "Real-time skeleton-tracking-based human action recognition using kinect data", In *MMM*, **1**, pp. 473–483 (2014).

33. Fisher, R.A. "The use of multiple measures in taxonomic problems", *Annals of Eugenics*, **7**, pp. 179–188 (1936).

34. Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edn., Springer, New York (2009).

35. Bishop, C., *Pattern Recognition and Machine Learning*, Springer, New York (2007).

36. Theodoridis, S. and Koutroumbas, K., *Pattern Recognition*, 2nd Edn., Elsevier Academic Press, USA (2003).

37. Rabiner, L.R. "A tutorial on hidden markov models and selected applications in speech recognition", In *Proceedings of the IEEE*, **77**(2), pp. 257–286 (1989).

38. Tst fall detection dataset. https://ieee-dataport.org/documents/tst-fall-detection-dataset-v2, Accessed:July 15, 2017.

39. Gasparrini, S., Cippitelli, E., Gambi, E., Spinsante, S., Wahsléen, J., Orhan, I., and Lindh, T. "Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion", In *ICT Innovations*, Springer, pp. 99–108 (2016).

40. Aggarwal, J.K. and Xia, L. "Human activity recognition from 3D data: A review", *Pattern Recognition Letters*, **48**, pp. 70–80 (2014).

41. Han, J., Shao, L., Xu, D., and Shotton, J. "Enhanced computer vision with microsoft kinect sensor: A review", *IEEE Transactions on Cybernetics*, **43**(5), pp. 1318–1334 (2013).

42. Chen, L., Wei, H., and Ferryman, J. "A survey of human motion analysis using depth imagery", *Pattern Recognition Letters*, **34**(15), pp. 1995–2006 (2013).

43. Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., and Gall, J. "A survey on human motion analysis from depth data", In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, Springer*, pp. 149–187 (2013).

44. Wang, J., Liu, Z., and Wu, Y. "Learning actionlet ensemble for 3d human action recognition", In *Human Action Recognition with Depth Cameras*, Springer, pp. 11–40 (2014).

45. Althloothi, S., Mahoor, M.H., Zhang, X., and Voyles, R.M. "Human activity recognition using multi-features and multiple kernel learning", *Pattern Recognition*, **47**(5), pp. 1800–1812 (2014).

46. Theodorakopoulos, I., Kastaniotis, D., Economou, G., and Fotopoulos, S. "Pose-based human action recognition via sparse representation in dissimilarity space", *Journal of Visual Communication and Image Representation*, **25**(1), pp. 12–23 (2014).

47. Kapsouras, I. and Nikolaidis, N. "Action recognition on motion capture data using a dynemes and forward differences representation", *Journal of Visual Communication and Image Representation*, **25**(6), pp. 1432–1445 (2014).

48. Liu, A.A., Nie, W.Z., Su, Y.T., Ma, L., Hao, T., and Yang, Z.X. "Coupled hidden conditional random fields for rgb-d human action recognition", *Signal Processing*, **112**, pp. 74–82 (2015).

49. Lee, S., Le, H.X., Ngo, H.Q., Kim, H.I., Han, M., Lee, Y.K., et al. "Semi-markov conditional random fields for accelerometer-based activity recognition", *Applied Intelligence*, **35**(2), pp. 226–241 (2011).

Biographies

Mozhgan Mokari received her BSc degree in Electrical Engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran in 2014. She also received her MSc degree in the same field from Sharif University of Technology, Tehran, Iran in 2016. She is currently studying for PhD in Electrical Engineering in Sharif University of Technology, Tehran, Iran. Her research interests are machine learning, computer vision, and signal processing.
ORCID IDs: https://orcid.org/0000-0002-1707-7907

Hoda Mohammadzade received her BSc degree from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran in 2004, the MSc degree from the University of Calgary, Canada in 2007, and the PhD degree from the University of Toronto, Canada in 2012, all in Electrical Engineering. She is currently an Assistant Professor of Electrical Engineering at Sharif University of Technology, Tehran, Iran. Her research interests include signal and image processing, computer vision, pattern recognition, biometric systems, and bioinformatics.
ORCID IDs: https://orcid.org/0000-0002-9852-5088

Benyamin Ghojogh obtained his first and second BSc degrees in Electrical Engineering (electronics and telecommunications) from Amirkabir University of Technology, Tehran, Iran in 2015 and 2017, respectively. He also received his MSc degree in Electrical Engineering (digital electronic systems) from Sharif University of Technology, Tehran, Iran in 2017. He is currently studying for PhD of Electrical and Computer Engineering in University of Waterloo, Canada. One of his honors is earning the second rank in the Electrical Engineering Olympiad of Iran in 2015. His research interests include machine learning and computer vision.
ORCID IDs: https://orcid.org/0000-0002-9617-291X